NKS-481
ISBN 978-87-7893-577-9

Principle Component Analysis as Applied to Qualitative Analysis of Mobile Measurement and Monitoring Data Sets (PRICOMOB)

M. Dowdall[1]
Tero Karhunen[2]
Ellinoora Vikman[2]
Mats Eriksson[3]
Gísli Jónsson[4]

[1]Norwegian Radiation Protection Authority, PO Box 55
N-1332, Østerås, Norway
[2]Radiation and Nuclear Safety Authority of Finland, Jokiniemenkuja 1
01370 Vantaa, Finland
[3]Linköping University, SE-581 83 Linköping, Sweden
[4]Icelandic Radiation Safety Authority, Raudararstigur 10
150 Reykjavik, Iceland.

January 2024

Abstract

Mobile measurement systems are the backbone of most responses to cases of orphan sources. Conducting mobile measurement surveys, irrespective of the platform utilised, is a non-trivial task with respect to the nature of the data being accrued – large volumes of discrete, often highly variable, data points where the signal of interest may be weak, superimposed on a constantly fluctuating background and only present for a tiny proportion of the overall data set. Principal Component Analysis (PCA), one of the most popular multivariate statistical technique, is a flexible statistical procedure that allows for the summarizing of the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed in order to observe trends, jumps, clusters and outliers. The PRICOMOB project focussed on the application of PCA to mobile measurement and stationary scanning data to assess its performance in identifying source signals from a number of isotopes superimposed on a variable background signal typical of mobile measurement data. The PCA method implemented proved itself a viable method to detect anomalies in spectral time series. A disadvantage of the method employed is that a training data set is needed containing all the features and behavior that are not due to artificial radioactivity. Alternative ways to form the residuals used in deciding whether a measurement contains features not previously seen included the Mahalanobis distance and a modified Euclidean distance. The modified Euclidean distance seemed to result in improved sensitivity for radionuclides that produce peaks, but reduced sensitivity for sources that produce continuums (such as x-rays).

Key words

Principal Component Analysis, Gamm spectrometry, Mobile measurement, time series

NKS-481
ISBN 978-87-7893-577-9
Electronic report, January 2024
NKS Secretariat
P.O. Box 49
DK - 4000 Roskilde, Denmark
Phone +45 4677 4041
www.nks.org
e-mail nks@nks.org

Nordisk kernesikkerhedsforskning
Norrænar kjarnöryggisrannsóknir
Pohjoismainen ydinturvallisuustutkimus
Nordisk kernesikkerhedsforskning
Nordisk kärnsäkerhetsforskning
Nordic nuclear safety research

December 2023

Principle Component Analysis as Applied to Qualitative Analysis of Mobile Measurement and Monitoring Data Sets (PRICOMOB)

Final Report from the NKS-B Project PRICOMB (Contract: AFT/B(23)1).

M. Dowdall[1], Tero Karhunen[2], Ellinoora Vikman[2], Mats Eriksson[3] Gísli Jónsson[4]

[1] Norwegian Radiation Protection Authority, PO Box 55, N-1332, Østerås, Norway

[2] Radiation and Nuclear Safety Authority of Finland, Jokiniemenkuja 1, 01370 Vantaa,Finland

[3] Linköping University, SE-581 83 Linköping, Sweden

[4]Icelandic Radiation Safety Authority, Raudararstigur 10, 150 Reykjavik, Iceland.

Table of Contents

1.0		**Introduction**	4
	1.1	Principal Component Analysis (PCA)	5
	1.2	Principal components and their geometric interpretation	7
2.0		**Methods**	10
	2.1	The Data Sets	10
	2.2	PCA Approach I	13
	2.3	PCA based approach for monitoring measurements.	14
	2.4	The FFM algorithm in more detail.	18
	2.5	The FFM algorithm summary	20
3.0		**Results and Discussion**	20
	3.1	PCA analysis of NaI dataset, PCA approach I	20
	3.2	Results on the FFM algorithm on monitoring data	23
4.0		**Conclusions**	28
5.0		References	29

DISCLAIMER

The mentioning of equipment, production companies, instruments, software or other tradenames does not constitute an endorsement on the part of NKS, participating teams or their institutions.

1. Introduction

Car borne and related (helicopter, drone, static scanning etc.) deployments of measurement systems are, and are likely to remain, a mainstay in the response arsenal of many countries to incidents involving searches devoted to the localization of gamma-ray sources. While in principle similar to laboratory-based systems, the operation of mobile measurement-based systems differs in a number of important ways. Mobile measurement typically generates significant amounts of spectral data where the individual measurements are of short duration (1 or 2 seconds each). Mobile measurement systems often deploy gamma detectors of relatively low resolution such as NaI or similar. The context within which such systems are deployed often necessitates rapid real-time analysis of this data or, alternatively, post -processing of the data at some later time. Mobile measurement, while well established and a mature technique, has undergone some changes in recent months/years. These are in relation to context – where mobile measurement is an invaluable assistance measure where another country requests help in finding or controlling orphan sources such as evidenced by the ongoing situation in Ukraine – and in technology – whereby new detector technologies (CdZnTe, LaBr) are being mounted on ever more flexible platforms (drones, etc.). These detector technologies are also employed in environmental monitoring applications, where similar analysis techniques are used as in the mobile applications. Concomitant with these developments has been an increased focus on analysis procedures and approaches such that the maximum benefit may be accrued from this measurement method and the difficulties inherent in it may be overcome.

Typically, mobile measurement and real-time monitoring systems have relied on data visualization systems involving the scrolling presentation of color-coded spectral data which, as the underlying principles are the same irrespective of specific implementations, are collectively referred to as “waterfall” displays (see Figure 1). While such systems function adequately for strong sources with emissions of some hundreds of keV and greater, displays of this type have some inherent disadvantages. Low level signals from artificial radionuclide sources can be difficult to separate from background variations, sources with emissions in the vicinity of strong background lines can be difficult to observe and gamma emitters where the emission is in the lower end of the energy spectrum can present difficulties. The display can be tiring to observe for long periods and is often unintuitive for inexperienced operators. Without other data handling procedures, the method is very operator dependent. Automatic identifiers of source signals, such as dose rates or count rates in “spectral windows” or Regions of Interest (ROI) are often vulnerable to Type I and Type II errors due to highly variable background signals upon which weak source signals may be superimposed or strong but highly localized sources of background due to geology etc.

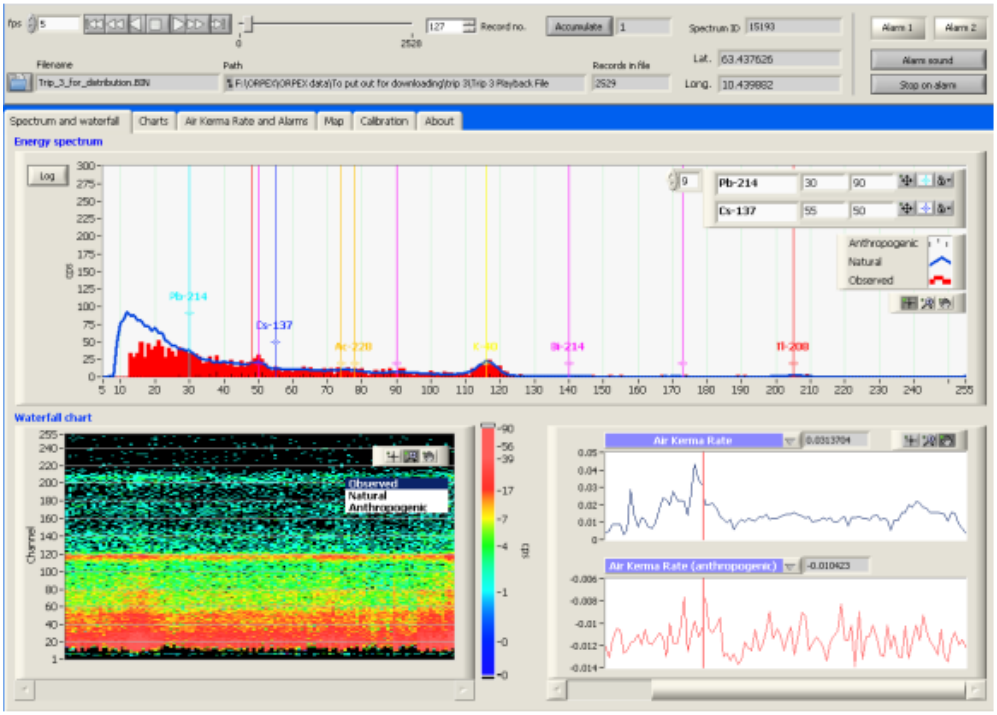

Figure 1. Conventional waterfall type display and ancilliary information depictions for mobile measurement systems.

1.1. Principal Component Analysis (PCA)

Principal Component Analysis, or PCA, is a general term for a set of statistical approaches that allows for the summarizing of the information content in large data sets by means of a smaller set of "summary indices" that can be more easily visualized and analyzed. Principal component analysis is one of the most popular multivariate statistical techniques. The most important use of PCA is to represent a multivariate data table as smaller set of variables in order to observe trends, jumps, clusters and outliers. PCA is a very flexible tool and allows analysis of datasets that may contain, for example, multicollinearity, missing values, categorical data, and imprecise measurements. The goal is to extract the important information from the data and to express this information as a set of summary indices called principal components or to "axis" in the multi-dimensional dataset where the data show the most deviations. Statistically, PCA finds lines, planes and hyper-planes in the K-dimensional space that approximate the data as well as possible in the least squares sense. A line or plane that is the least squares approximation of a set of data points makes the variance of the coordinates on the line or plane as large as possible.

Principal Component Analysis potentially offers several distinct advantages across a number of different fields including data analysis, machine learning and signal processing. PCA functions by transforming high-dimensional data into a lower-dimensional space which captures most of the

variability in the original data set. This reduction simplifies the dataset to some extent and can be advantageous in employing more efficient and faster computations. By identifying the most pertinent features of the data by focusing on the directions (principal components) in which the data set varies to the greatest extent, key information is retained while less important or redundant features of the data set are discarded or ignored. The reduced-dimensional representation derived by PCA facilitates easy visualization enabling the elucidation and understanding of patterns, relationships and clusters within the data set being analyzed

As PCA focuses on capturing only the most significant sources of variability within the data, less important variations, which may be considered as noise, may be minimized or ignored ultimately producing a cleaner, more robust representation of the patterns inherent in the data.

For datasets where variables exhibit high degrees of correlation, PCA can be of utility in decorrelating these variables which is of some value in regression analysis. The principal components obtained by PCA are uncorrelated, simplifying the interpretation of the transformed data and facilitating subsequent analyses. For large data sets where computational overheads may be onerous, producing and working with a lower-dimensional representation of the data can, for example, significantly speed up the training of machine learning algorithms and reduce the overhead. PCA finds applications in a variety of fields due to its versatility. These include machine learning, image analysis and processing, bioinformatics, genomics, economics and finance, chemometrics, spectroscopy, signals processing, geophysics, psychology and social sciences, marketing, facial recognition, remote sensing and medical imaging. Introductions to the theory and practice of PCA may be found in Jolliffe (2014) and James et al. (2014).

PCA has been applied previously to certain aspects of the general gamma ray spectrometry field. Kishimoto et al. (2021) recently applied PCA to the optimization of search strategies for robot borne detectors, Pires de Lima and Marfurt (2018) having previously used PCA for analysis of natural gamma signatures in airborne spectrometry. Reinhardt (2014) applied PCA to analysis of NaI(Tl) spectra, suggesting it as means to cope with time-varying background, and using it as a smoothener of the statistical fluctuation of channel contents. Williams (2019) developed PCA methods for source localization using directional CdZnTe detectors while Minty and Hovgaard (2002) utilized PCA for reduction of noise in gamma ray spectra. Of most relevance to this proposal, PCA analysis has been used as a means of detection of anomalous spectra at radiation portal monitors – outperforming commercial solutions and functioning well with low count rate spectra (Boardman et al., 2012). Eriksson and Dowdall (2021) recently demonstrated the application of PCA to categorization of spectral data sets in relation to special nuclear materials.

The PRICOMOB project focused primarily on the application of PCA to mobile measurement data to assess its performance in identifying source signals from a number of isotopes superimposed on a variable background signal typical of mobile measurement data. The application of PCA to this field brings a powerful statistical technique to bear on data sets that are, relative to laboratory based gamma ray spectrometry, somewhat complex and that present challenges on a number of fronts.

1.2. Principal components and their geometric interpretation

For a set of n variables X_n, the k:th principal component

$$PC_k = a_{k1}X_1 + a_{k2}X_2 + \cdots + a_{kn}X_n$$

is a new variable formed as a linear combination of the original variables in such a way, that the first principal component contains the largest amount of variation in the data, the second principal component the second largest and so on. Additionally, the principal components will be uncorrelated.

The above definition and formula are not very intuitive. A geometric interpretation is given and illustrated in two dimensions in the following.

Consider a set of two dimensional datapoints drawn from a multivariate normal distribution. The distribution of the datapoints is characterized by mean vector μ and covariance matrix Σ (Figure 2).

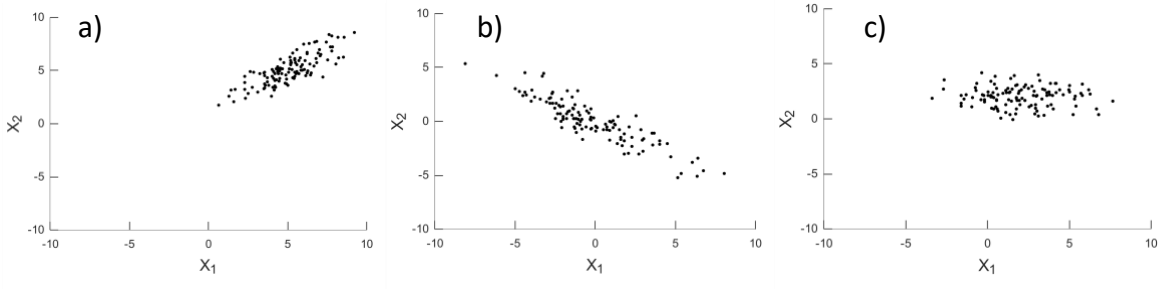

Figure 2. Data points drawn from a two-dimensional multivariate normal distribution. The underlying distributions are characterised in a) $\mu = [5, 5]$, $\Sigma = [3\ 2; 2\ 3]$ in b) $\mu = [0, 0]$, $\Sigma = [8 - 5; -5\ \ 4]$ and in c) $\mu = [0, 0]$, $\Sigma = [5\ 0; 0\ 5]$.

For this two-dimensional case, the probability density function (PDF) of the underlaying distribution can be plotted (Figure 3).

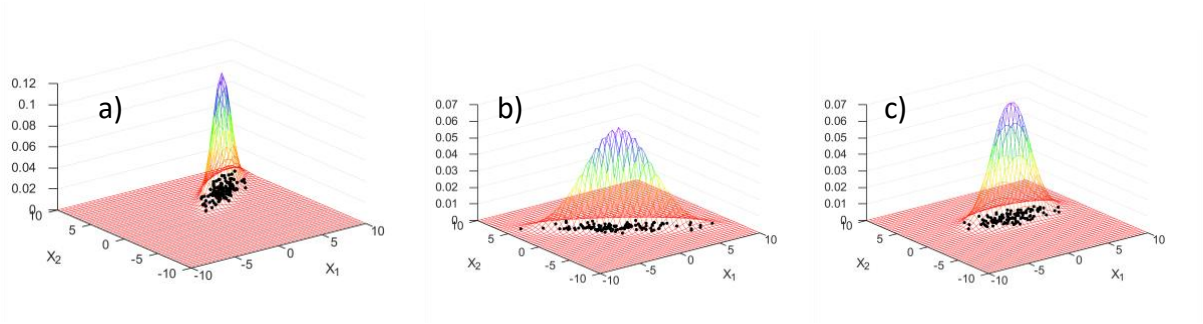

Figure 3. The data points from Figure 2 and the PDF of their underlying distribution.

The location of the hump of the PDF is determined by the mean vector μ. The width and rotation of the hump is determined by the covariance matrix Σ. Contours of constant density can also be plotted, and for the multivariate normal distribution they will be ellipses. The principal components correspond to the principal axes of these ellipses. The principal axes are in the direction of the eigenvectors of the covariance matrix (Figure 4). The eigenvector corresponding to the largest eigenvalue is the first principal component and corresponds to a principal axis of the ellipsoid containing the data points.

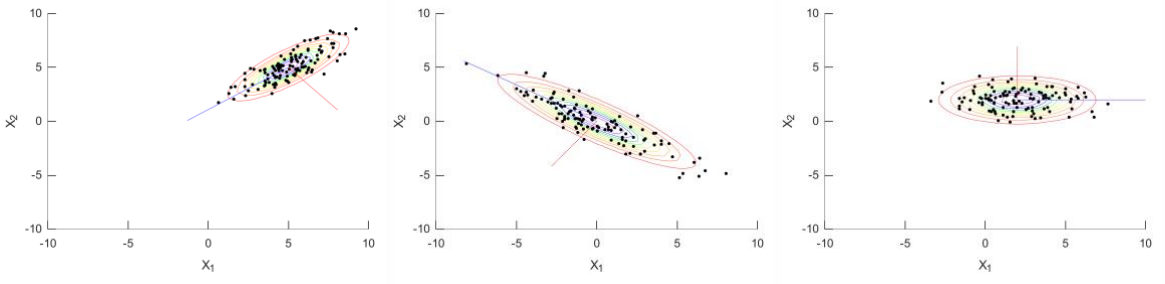

Figure 4. Contours of constant density (ellipses) and the eigenvectors of the covariance matrix for the data points of Figure 2.

Thus, to determine the principal components for a given set of datapoints, it becomes necessary to estimate the mean vector and the covariance matrix of the distribution the data points were drawn from. The eigenvectors – and thereby the principal components - can then be obtained from the estimated sample covariance matrix. The principal components can be arranged into a matrix V, that can be used to project the datapoints onto the principal components. By forming the matrix $B = V^T V$, the projection matrix, the datapoints represented by the principal components are projected back to the original basis to reconstruct the data. The dimensionality of the data is reduced by using only few of the first principal components in the construction of V (Figure).

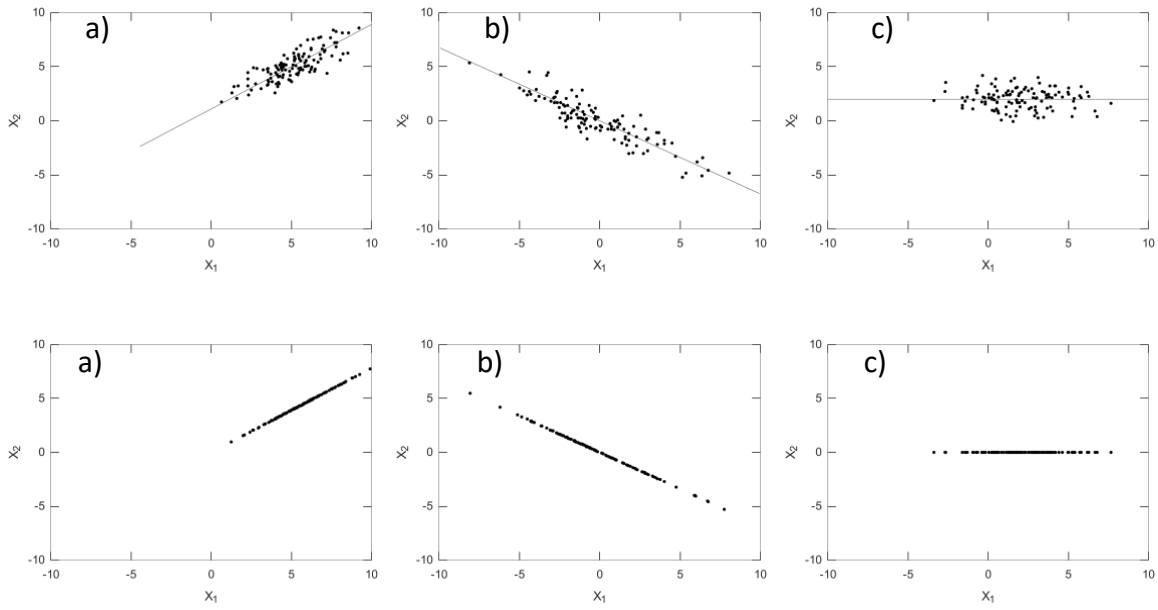

Figure 5. Top row: The data points of Figure 2 plotted with the principal axis corresponding to the eigenvector with the largest eigenvalue. Bottom row: The reconstruction of the datapoints in terms of the largest principal component. The dimensionality has been reduced from two to one.

A reconstruction similar to Figure 5 (b) can be made with GNU Octave with the following code:

```
pkg load statistics;
mu = [0 0];
sigma = [8 -5;-5 4];
x = mvnrnd(mu, sigma, 120);
[V l] = eigs(sigma);
B = V(:,1) * V(:,1)'
pr = (B*x')';
plot(pr(:,1),pr(:,2), "k.");
```

The above intuition holds for dimensions greater than two, in which case the contours of constant probability are ellipsoids of appropriate dimension. The principal components then correspond to the principal axii of the ellipsoids. In practice, the number of dimensions is large (e.g. 512 in the algorithm of Section 2.3), and the reduction in dimensionality of data is done to facilitate further calculations.

The number of principal components to select for representing the data can be made by considering the portion of total variance of the data that they explain. This is called the cumulative explained variance. Details on the theory of PCA relevant to spectrum analysis in the scope of this project are given in (Vikman 2023).

2. Methods

2.1. The Data Sets

A series of data sets were generated for the purposes of PRICOMOB. These can be conveniently divided into those based around an LaBr detector system and an NaI detector system. The data sets were based upon actual data into which synthetic data was then inserted. For the LaBr system, the base data was from a static scanning system and for the NaI base data, the data was obtained from a mobile measurement system. The LaBr detector was a 1.5" standard detector and the NaI detector was a 4 l standard detector. The LaBr detector accrued data over 2048 channels for 600 s interval with the NaI detector accruing data every second over 1024 channels. For the NaI data set, point sources of Xe-133, Cs-134, Ir-192, Am-241, Ra-226, Co-60, I-131 and Cs-137 were simulated as passing the detector at distances of 4 m, 8 m, 12 m, 16 m, 20 m, 24 m and 28 m resulting in steadily weaker signals. The passage of the source in front of the detector was over a time of 10 s. The simulated data was then added to the actual data. For the LaBr data set, point sources of Ba-133, Xe-133, Cs-134, I-131, Am-241, Co-60 and Cs-137 were simulated as passing in front of the detector at distances of 40 cm, 120 cm, 200 cm, 280 cm, 360 cm and 480 cm resulting in decreasing signals with distance.

Figure 6. Mobile track, from south Svanvik to north Kirkenes (blue line). Red dot in the middle of the track indicate the radioactive source postion. Map produced by ustin the online service GPS Visualizer tool (https://www.gpsvisualizer.com/map_input).

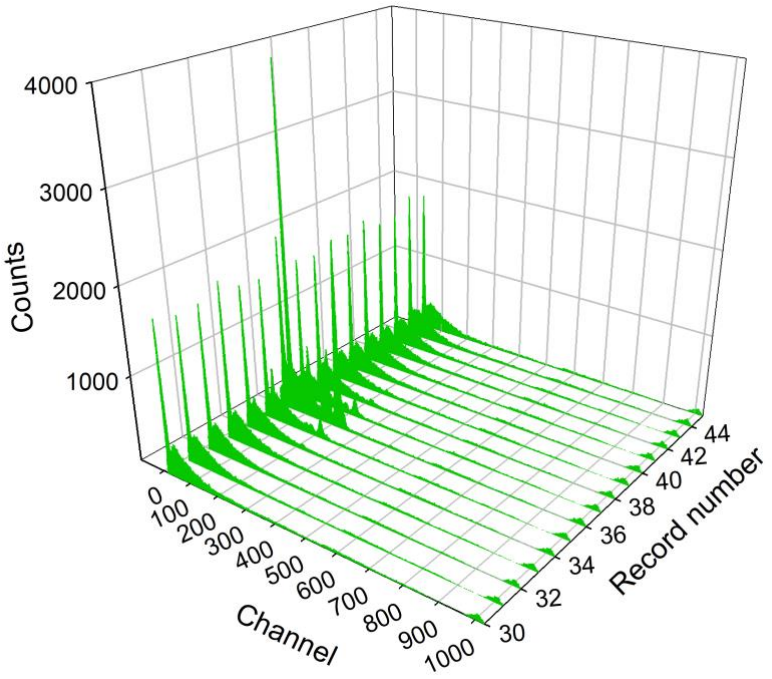

Figure 7. Selection of spectral records from the LaBr base data set (channels 0 to 1000) with inserted Ba-133.

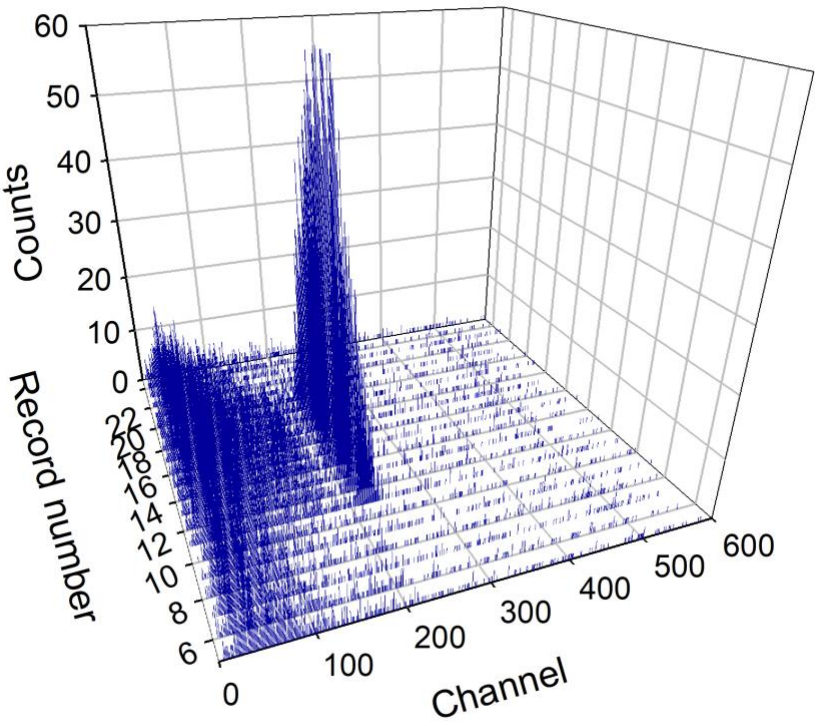

Figure 8. Selection of spectral records from the NaI base data set (channels 0 to 600) with inserted Cs-137.

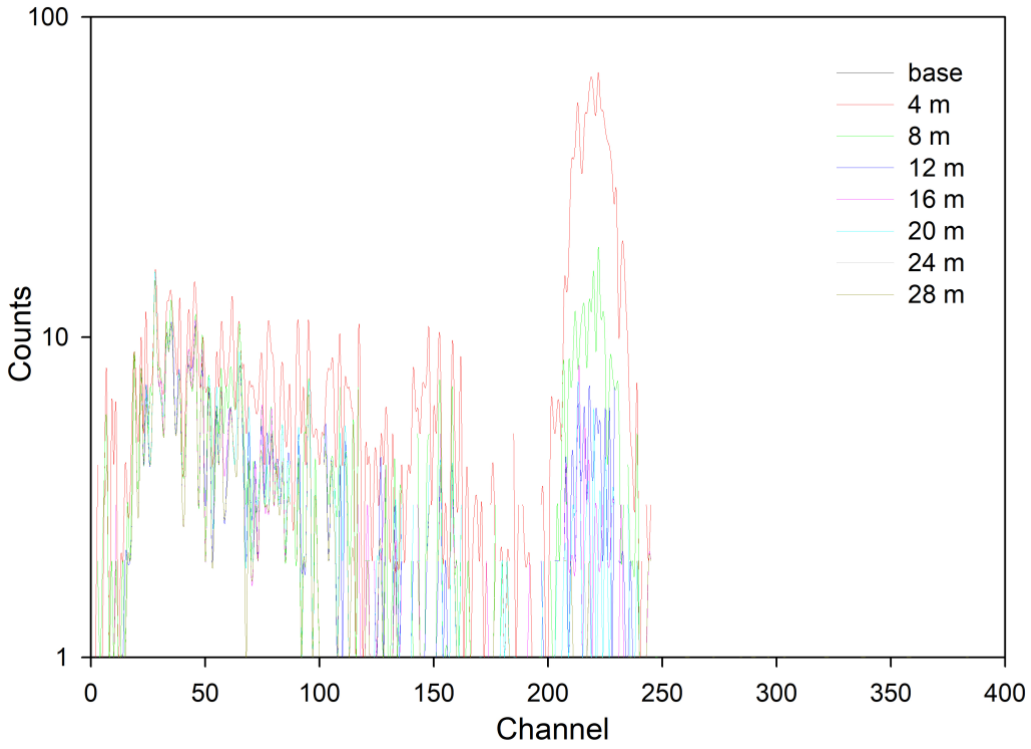

Figure 9. Cs-137 signals in one channel as function of distance between channel 0 and 400 for the NaI detector.

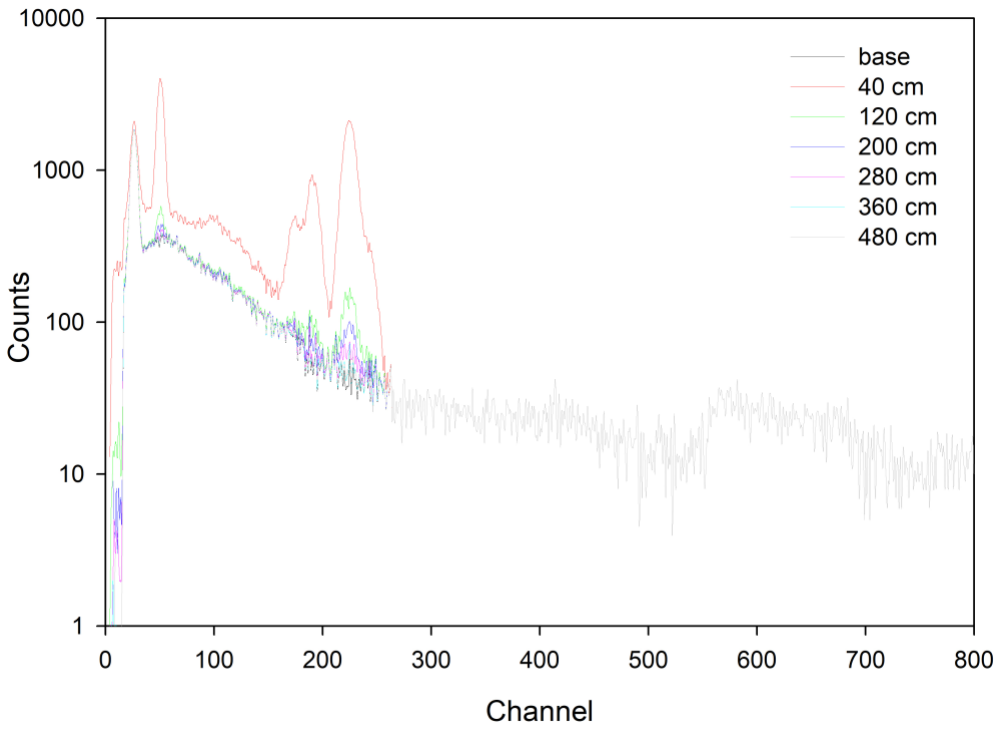

Figure 10. Ba-133 signals in one channel as function of distance between channel 0 and 400 for the NaI detector.

2.2 PCA approach I

To apply the PCA theory into mobile gamma spectrometry it can be simplified as, the entire dataset to be analyzed can be seen as all the spectra recorded during one driven path, e.g. see Figure 6. Each channel in the spectra will be a variable of the dataset. The dataset can easily add more variables after the last energy channel, e.g., ROI:s of radionuclides, sum of all channels in the low energy region and high energy region, altitude, etc. It is of course also possible to exclude a particular part of the spectrum, e.g., noisy low energy channels where no interesting information might be recorded. The dataset must be normalized, i.e., this can be done by all channel's being divided by the maximum number of counts in the spectra or normalized with live time, i.e. use counts per second instead of counts. It is important to normalize all spectra in the dataset the same way otherwise a biased PCA analysis will occur.

In this approach we have used the NaI detector dataset described above with the inserted signals and these have been analyzed with the open-source program R Statistical Software (v4.3.1; R Core Team 2021). The dataset contains 2678 recorded spectra with 1024 channels each. After the last energy channel, we have added the sum of counts in region of interest (ROI) for each radionuclide peak energy. For radionuclides emitting multiple gamma energies ROI over each energy was summarized to one ROI. The radionuclide ROI was background corrected by subtracting the sum of counts from similar ROIs above and below the radionuclide peak ROI. In total we added 9 ROI (in addition to the 8 radionuclides listed in section 2.1. we have added ^{40}K) variables into the dataset. First, we normalized the full dataset with the maxima count in the dataset, after this the ROI counts were analyzed and added to the dataset. These background corrected ROIs were normalized as, a) if the value were below 1 the ROI was set to 0 and b) for values higher than 1 the ROI number where multiplied a withing factor of 2 to enhance a possible detection of sources along the given path.

In this approach we read in the simulated dataset, a, into R by the function "read.csv", normalized the dataset "a<-a/max(a)". After this the ROIs where calculated as described below:

```
Cs137ch<-(221-10):(221+10) # 137Cs peak ROI range in channels, 3Kev/ch, ch: 201-231

HBkgCs137ch<-232:252 # 137Cs background ROI range in channels, high energy side

LBkgCs137ch<-190:210 # 137Cs background ROI range in channels, high energy side

Cs137ROI<-c(1:sp_n) # dummy vector to record all ROI for all spectra in the data set, sp_n=2678

for (ii in 1:sp_n)Cs137ROI[ii]<-sum(a[ii,Cs137ch])-
(sum(a[ii,LBkgCs137ch])+sum(a[ii,HBkgCs137ch]))/2 # calculating net counts in ROI, tot-bkg
```

```
Cs137ROI[Cs137ROI<1]<-0 # if ROI value is below 1 its given a value 0
Cs137ROI<-2*Cs137ROI # the ROI is multiplied by a weighting factor of 2
```

After all ROIs have been calculated they are added to the dataset by the "cbind()" function, e.g. a<-cbind(a,Cs137ROI). Finaly the PCA analysis of the dataset is performed with the "prcomp()" function, e.g. PCA<-prcomp(a). The results of the PCA analysis can be represented in may different ways, but we have choose to use the "biplot()" function that describes the results in an intuitive manner to understand the variations in the dataset. We used this analyzing approach for all the simulated datasets.

2.3 PCA based approach for monitoring measurements.

A PCA based approach was studied for analysis of monitoring data, i.e., for data produced by stationary spectrometric measurements of the environment. The objective was to detect spectra containing peaks (or other responses) due to artificial radioactivity or radiation (Figure 11).

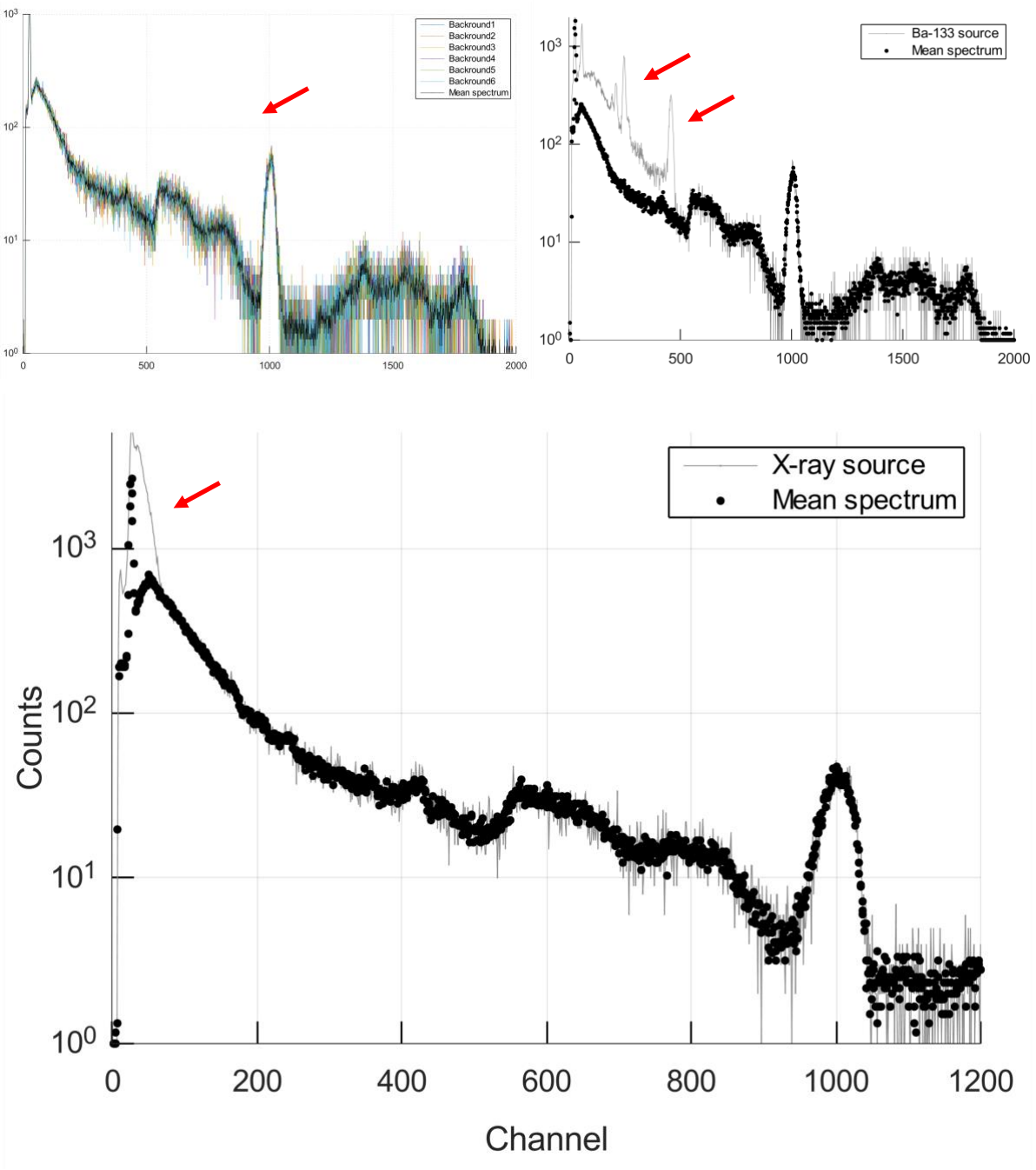

Figure 11. Top left: Typical $LaBr_3(Ce)$ background spectra and their mean spectrum. Top right: Mean spectrum of background and a measurement of a Ba-133 radionuclide source. Bottom: Mean spectrum of background and a measurement of an x-ray source.

The challenges in stationary monitoring arise from the effect of changing environmental conditions on the radiation situation. Such changes are, for example, rain events that flush airborne radon progeny down closer to the earth – and thereby closer to the detector. This causes an increase in the ambient dose rate – and the counting rate - at the detector location (Figure 12).

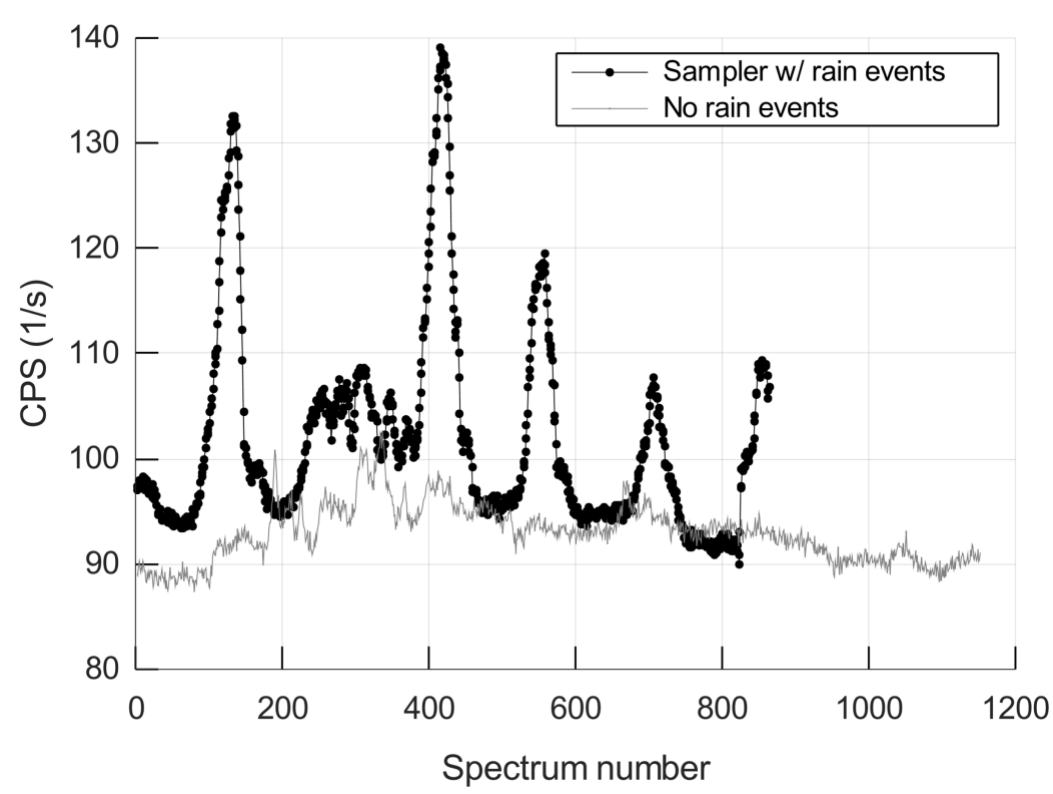

Figure 12. Changes in the counting rate during rain events. The rain events show as peaks in the timeseries of the total count rate of the measured spectra, contrasted to another time series from a period of more stable environmental conditions. Each spectrum is acquired with a 10 minute integration time.

The peaks corresponding to gamma emissions from radionuclides in the U-238 and Th-232 chains are also increased and become visible in the measured gamma spectra (Figure). These peaks can be misinterpreted as peaks corresponding to emissions from artificial radioactivity. An analysis algorithm should be able to adapt to this changing background somehow, and in so doing, be robust in the sense of not producing excessive false alarms due to environmental conditions.

A review of literature provided promising candidate method for this purpose. The general method described in (Jolliffe 2014) and the method described in (Boardman et al. 2012) for NaI(Tl) provided the starting point for implementation. A similar method, also based on the reconstruction using PCs, is presented in (Reinhardt 2014).

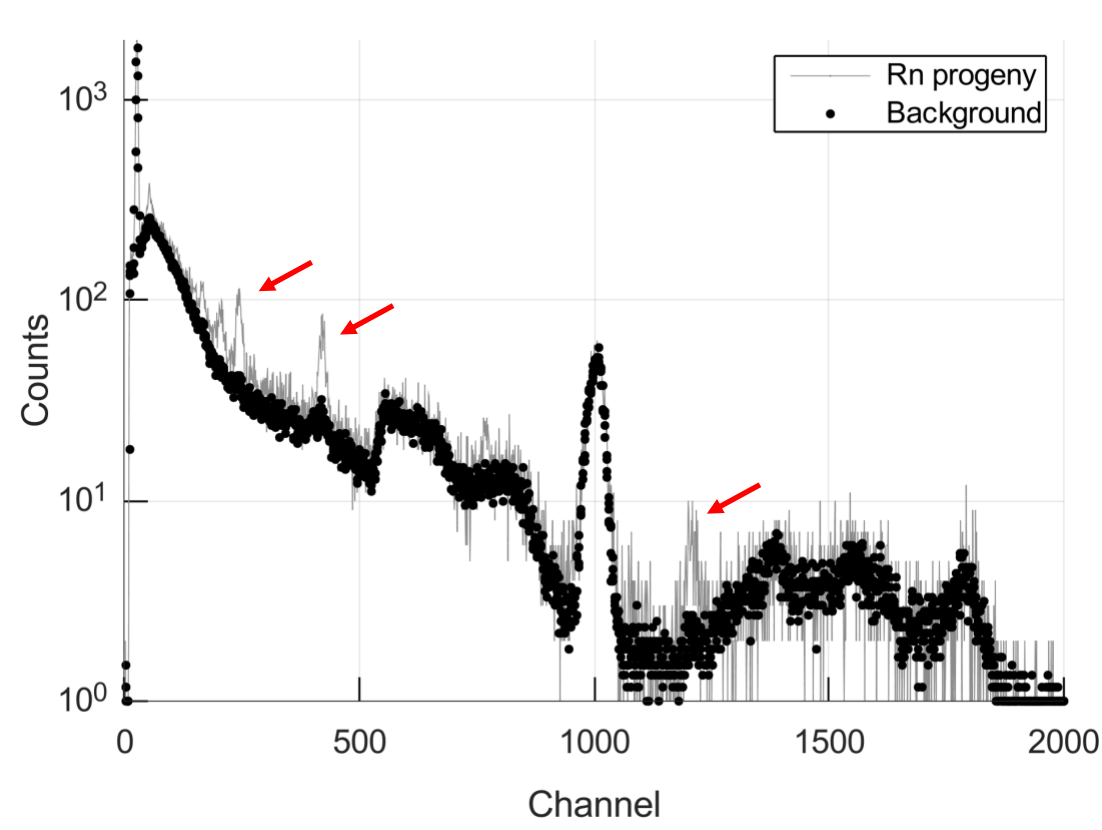

Figure 13. Peaks corresponding to gamma emissions from Radon progeny (grey) are visible during a rain event. The mean spectrum from background measurements (black) is from a period of no rain.

A variant of the algorithm was implemented, as described in (Vikman 2023). The resulting algorithm is based on a so-called Fixed Feature Model (FFM). A FFM is a form of feature learning model, where the fixed features are extracted from a given data matrix as its principal components. The data matrix in this case is formed from a set of spectra of background measurements called the *training data set*.

To obtain an alarm level, the statistical behavior of the extracted fixed features is examined by applying it on a set of spectra of background measurements called the *validation data set*.

The spectra to be analysed are called the *analysis data set* and are analyzed for the presence of artificial radioactivity using a reconstruction based on the fixed features (principal components) obtained from the training data set. The intuition for this reconstruction is given in Section 1.2, and the details are found in (Vikman 2023). If the reconstruction based on these features does not explain the spectrum, then it is concluded that artificial radioactivity is present.

More detail on how to form the FFM from the datasets (Figure 14) is given in the following section.

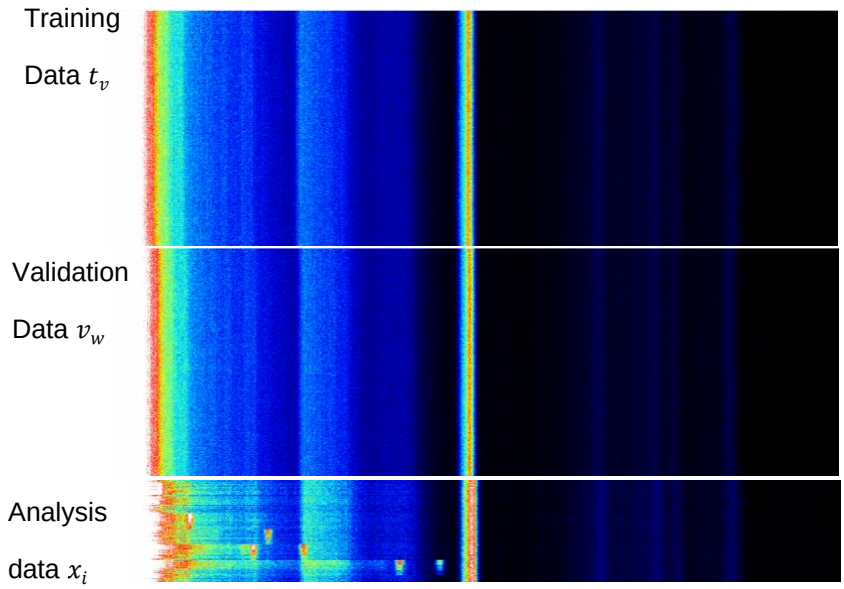

Figure 14. Waterfall plot of the training, validation and analysis data sets. The training and validation data should contain only background measurements that exhibit natural variation. The analysis data set is then analysed for signs of artificial radioactivity

2.4 The FFM algorithm in more detail.

The spectral data of the environmental measurements are interpreted as data points. A dataset thus consists of a 2048 dimensional vectors (datapoints) for each measurement.

The channel contents of each measurement are known to be Poisson distributed, suggesting that the multivariate normal distribution is a good representation of the underlaying distribution of the data points, provided that the number of counts in each channel is large enough to satisfy the Poisson – Normal approximation.

The lowest channels of the spectra will usually exhibit noise. So only channels above a minimum channel are considered. To further ensure that the data points are multivariate normal distributed, the spectra are compacted by summing together two consecutive channel contents. This is done twice, resulting in 512 channel spectra. The resulting summed channel contents will be Poisson distributed, with summed intensity of the original channels.

A scaling factor α is determined using the method of least squares for each spectrum in the training, validation, and analysis data sets. A mean spectrum of the training data is first calculated. The i:th channel of the mean spectrum is the mean of the i:th channels of the n spectra in the training data set

$$\mu_i = \frac{1}{n}\sum_i c_i.$$

The scaling factor α_j for the j:th spectrum is then

$$\alpha_j = \frac{1}{2}\sum_i \frac{c_i \mu_i}{c_i^2},$$

where c_i is counts in the i:th channel of the spectrum, and μ_i is the counts in the i:th channel of the mean spectrum as given above. By scaling each channel of the j:the spectra with the scaling factor α_j most spectra will be nearly identical.

The compacted and scaled spectra are standardized by subtracting the mean of the training data set and scaling with the standard deviation of the training data for each channel of each spectrum. The spectra are now preprocessed for further use. The principal components are then extracted from the training data set using the singular value decomposition. A data matrix M is formed, with the preprocessed training data points as its rows. The singular value decomposition for the data matrix M is

$$M = ULA^T.$$

The matrix of right singular vectors A contains the principal components of the training data as its columns. The principal components will capture the features contributing to the total variance of the training data set. The validation data set is then analyzed using the fixed features by comparing the difference of a data point and its reconstruction using the selected number of principal components. The selected number of principal components is used to form the columns of the matrix V. The reconstruction is made by forming the projection matrix $B = V^T V$. The difference between the original data point x and its reconstruction Bx is then the Euclidean distance

$$\|Bx - x\|,$$

and is here called the *residual*.

It is shown in (Vikman 2023) that the theoretical distribution of the residuals for Hermitian matrices B and random vectors x is a shifted Tracy – Windom distribution. The Tracy – Widom distribution can be approximated accurately using a shifted gamma distribution. By fitting the theoretical distribution onto a histogram of the residuals of the validation data, an alarm level can be established by finding the p-value corresponding to the prescribed false alarm probability.

When the FFM is then applied on the analysis data, an alarm is raised if the alarm level is exceeded by a residual. This signals that the spectrum being analyzed contains something else than the fixed features extracted from the training data.

2.5 The FFM algorithm summary

The algorithm can be summarized as

1. Preprocess the training (t_v), validation (v_w) and analysis (x_i) data sets
    a. Remove channels under prescribed minimum channel.
    b. Compact (summation of adjacent channels) the spectra two times.
    c. Apply LSQ scaling to the spectra.
    d. Standardize the spectra (subtract mean and scale by standard deviation).

2. Determine PCs from the training data (t_v) using SVD.
    a. Form the data matrix M from the preprocessed vectors t_v.
    b. Form V, the principal component matrix from a selected number of right singular vectors of M.
    c. Form $B = V^T V$, the projection matrix.

3. Determine the alarm level from the validation data.
    a. Represent validation data using PCs (calculate Bv_w for each validation data point v_w).
    b. Histogram the Euclidean distances $\|Bv_w - v_w\|$
    c. Fit the gamma distribution to the histogram.
    d. Find the p-value corresponding to given probability of false positives.

4. Analyze the spectra in the analysis data
    a. Represent validation data using PCs (calculate Bx_i for each analysis data point x_i).
    b. Calculate the Euclidean distances $\|Bx_i - x_i\|$.
    c. If the distance $\|Bx_i - x_i\|$ exceeds alarm level, raise alarm.

For each step, the details are given in (Vikman 2023).

The source code of a java language implementation of the FFM algorithm is found in an appendix of (Vikman 2023), and will be made available at https://github.com/StukFi.

3 Results and Discussion

3.1 PCA analysis of NaI dataset, PCA approach I

There are many ways to represent PCA results, and we present a way that we believe are intuitive and easy to use and understand. We are searching/looking for spectrum in the dataset that show some peculiar features not easy to detect by other means than PCA analysis, i.e. a “small” peak(s) origin from a source present along the path of the mobile track. In figures 15 and 16 an example of representation

of the PCA analysis is given. In the appendix it can be found a complementary set of figures showing the PCA analysis for each simulated set, focused on the last detectable and also given example when the PCA code used in this project could not identify any source along the path. In table 1 the results are summarized for all datasets. In our analysis all cases with radionuclides placed up to 16 m from the road could be detected, and for the ^{134}Cs source detection out to 28 m was possible, see Figure 15.

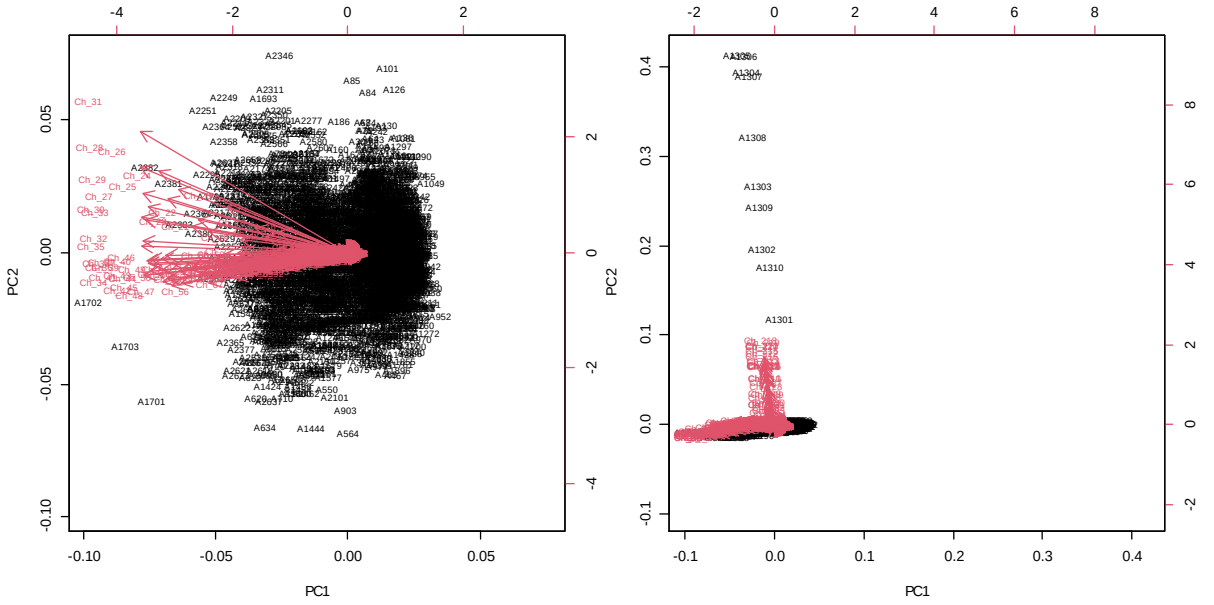

Figure 15. Example of PCA analysis using the biplot() function, black text gives the A followed with the spectra, rea number and arows gives the channels showing largest variation. To the left the mobile path (see figure 1) without any source, i.e. showing the variation of the background counts. To the right, the same path with a ^{137}Cs placed 4 m from the road. It is easily seen which spectra where we have ^{137}Cs (records A1301 to A1310). These analyses were done without any ROI applied.

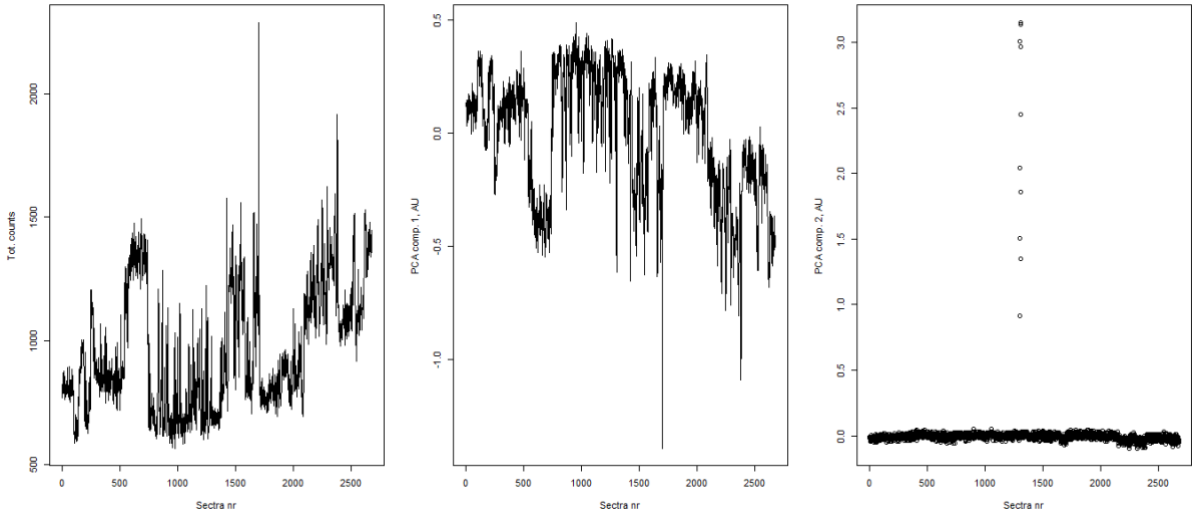

Figure 16. The ^{137}Cs data set with the source 4 m from the road. To the left the total counts in each spectrum along the track showing how the background count are changing with factor of more than 3. Middle the PCA component 1, i.e. the component explaining most of the variation in the data set, in this case clear inverse relationship between background and PCA comp 1 can be seen and its not easy to detect the ^{137}Cs source. To the right the PCA component 2 that show the deviation in the spectra A1301 to A1310 clearly the presence of the source.

NaI	4m	8 m	12 m	16 m	20 m	24 m	28 m
Cs-137							
Co-60							
I-131							
Am-241							
Ra-226							
Ir-192							
Cs-134							
Xe-133							

Table 1. PCA analysis of source identification using NaI detector. Table contains the 56 (8 different radionuclides at 7 distances datasets simulated. Green means that the PCA analysis found the right radionuclide and identified in which spectra detections was done. Light green indicates that a week signal could be detected but in a clear identification in which spectra. White square means that the PCA in this setup could not be detected.

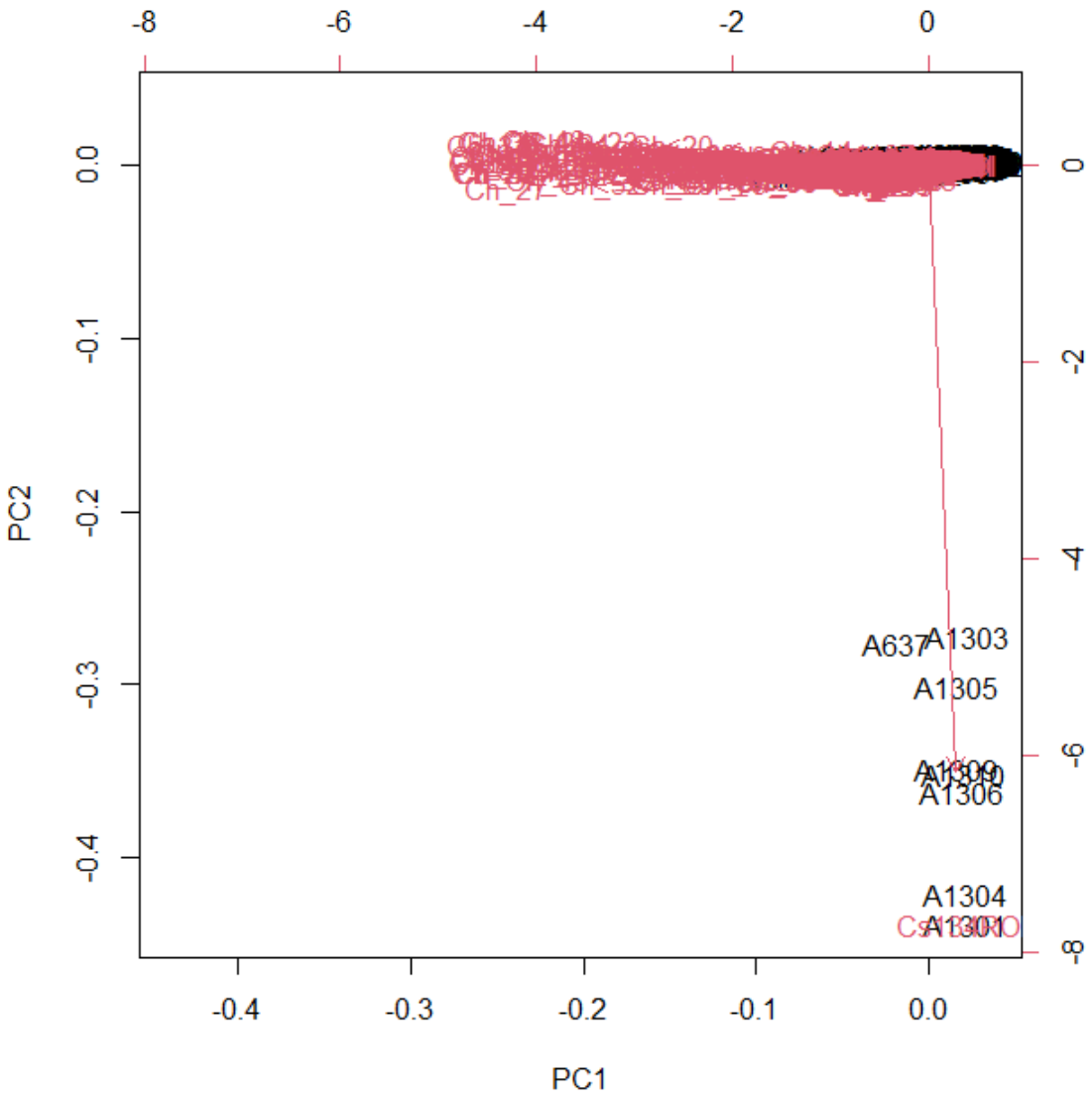

Figure 17. Biplot of the PCA analysis of the dataset containing a ^{134}Cs source placed 28 m from the road. The spectra A637 is a false positive result, however the records A1301-A1309 are correct identified.

3.2 Results on the FFM algorithm on monitoring data

The FFM algorithm of section 2.3 was tested by Vikman (2023) on monitoring data from two spectrometric dose rate monitoring stations – Nuorgam and Rovaniemi. The Rovaniemi station is an air sampling station, and the spectrometer monitors the sample while its being collected. The Nuorgam station is an *in-situ* spectrometer making direct measurements of the environment. The spectrometer used at both stations is a 1.5" $LaBr_3$(Ce) scintillator. Both spectrometers produce 2048 channel pulse height distributions. The integration time of the measurements is 10 minutes.

The data from these stations were used *as is* to form the training data set and the validation data set. The analysis data set was formed from another portion of the measurement data by adding synthetic spectral responses of selected radionuclides to the measurements. The data set sizes used are given in Table 2. The parameters for the algorithm used in the testing are given in Table 3. The number of principal components to use in the reconstruction was selected based on the cumulative explained variance.

Station	**Training data size**	**Validation data size**	**Analysis data size**
Rovaniemi	19 716	8 451	28 168
Nuorgam	19 841	8504	28 346

Table 2. Data set sizes for the testing of the FFM algorithm.

Parameter	**value**	**Comment**
Number of compactions	2	Results in a 512 channel spectrum
Number of principal components	9	Increase in cumulative explained variance is negligible beyond 9 of the largest PCs (**Error! Reference source not found.**).
False alarm probability	7.7e-6	Corresponds to one false alarm per month from 30 stations.
Minimum channel	44	Corresponds to energy of ~46 keV.

Table 3. Parameters used in the testing of the FFM algorithm.

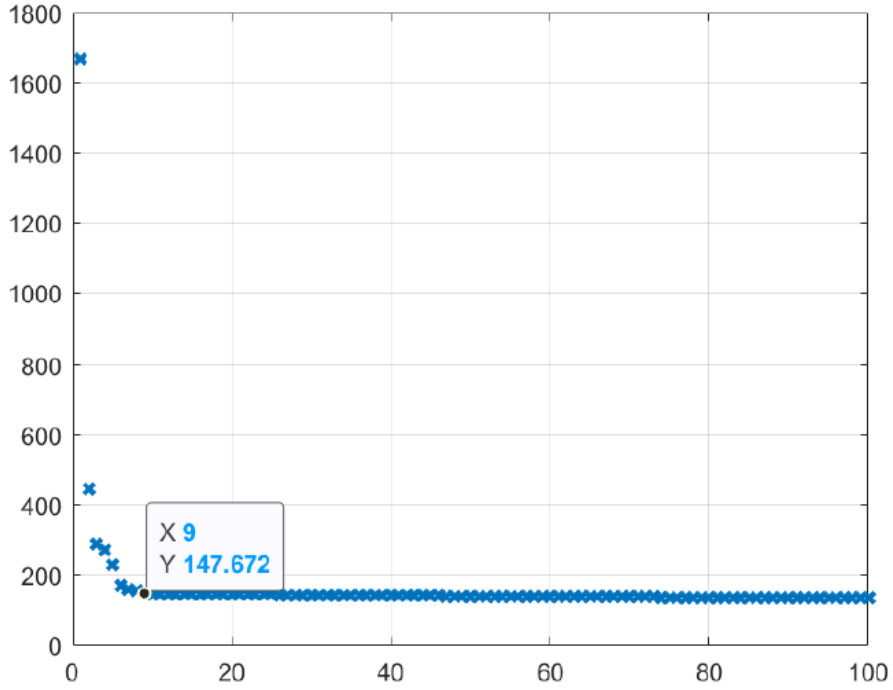

Figure 18. Variance of the PCs as the function of the PC number for the Nuorgam validation data. Figure from (Vikman 2023).

The activities for the synthetic spectral responses added to the validation data set are given in Table 4. Ten responses of each activity were added to background measurements, to study the effect of the background on the detection.

The synthetic responses were generated by adding Gaussian peaks and Compton continuums corresponding to radionuclide emissions onto the spectra. The Gaussian peaks were generated based on a reference area specified for one peak corresponding to given emission. The sizes of the other peaks of the response are then calculated based on nuclide data – gamma and x-ray yields – and the efficiency calibration. The efficiency calibration used in generating the spectral responses is the computational calibration presented in Dowdall et. al (2022). The Compton continuum shape is formed using the Klein – Nishina equation as described in Kudomi (1999). The area of the continuum was calculated using the reference peak area and an empirically determined peak-to-compton ratio.

For both the Rovaniemi and Nuorgam data, the Tracy – Widom distribution is an excellent fit for the histograms of the residuals (**Error! Reference source not found.** 19). It is therefore believed that the prescribed false alarm rate is realized in the tests. Synthetically generated responses of the test radionuclides could be detected from the analysis data sets of both stations. The results are shown in Tables 4 and 5.

Table 4. The activities and detections of the synthetic spectral responses for the Rovaniemi data.

Radionuclide	Activity (Bq / m^3)	Detections (out of 10)
Am-241	50.2	0
	100.4	2
	150.6	7
Co-60	30.0	0
	60.0	8
	90.0	10
Cs-134	20.9	0
	41.9	1
	62.8	10
Cs-137	25.3	0
	50.6	0
	76.0	4
I-131	18.1	0
	36.2	0
	54.4	4
Xe-133	40.0	0
	80.0	0
	120.0	9
Xe-135	13.7	0
	27.5	1
	41.2	8

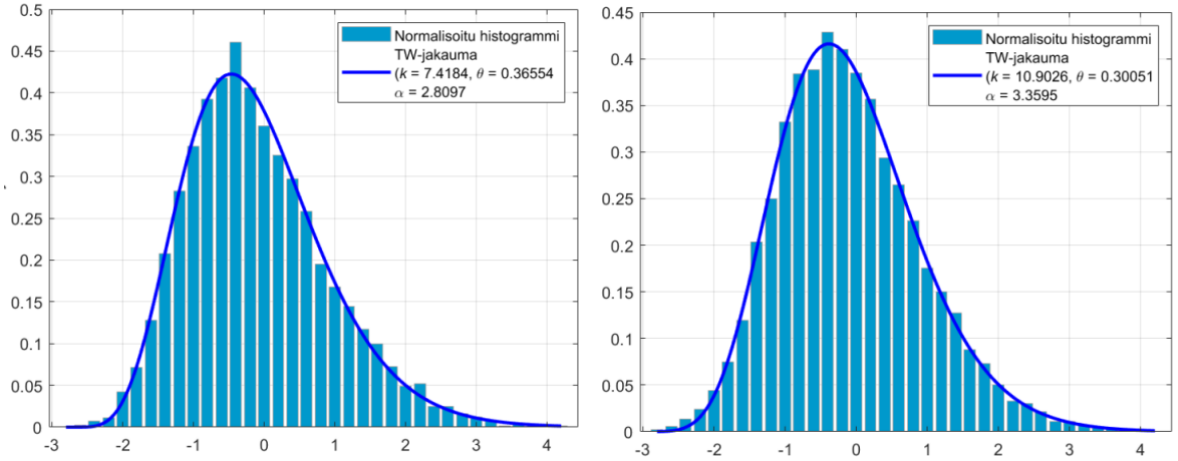

Figure 19. Left: Approximation of the Tracy – Widom (TW) distribution fitted to the histogram of standardised residuals of Nuorgam validation data. Right: TW approximation fitted to the histrogram of standardised residuals of Rovaniemi validation data. Figures from (Vikman 2023).

Table 5. The activities and detections of the synthetic spectral responses for the Nuorgam data.

Radionuclide	**Activity (Bq / m^3)**	**Detections (out of 10)**
Am-241	49.4	0
	98.8	0
	148.2	6
Co-60	29.5	0
	59.0	9
	88.5	10
Cs-134	20.6	0
	41.2	0
	61.8	5
Cs-137	24.9	0
	49.8	0
	74.7	6
I-131	17.8	0
	35.7	0
	53.5	1
Xe-133	39.3	0
	78.7	2
	118.1	3
Xe-135	13.5	0
	27.0	0
	40.5	3

The activities of the radionuclides are quite close to the detection limits of the method for the prescribed false alarm rate, as can be seen from the fact that the detections increase from 0 / 0 at the lowest activity level to almost 10/10 at the highest activity level. In the tests both datasets caused no false positive alarms (Figure 20).

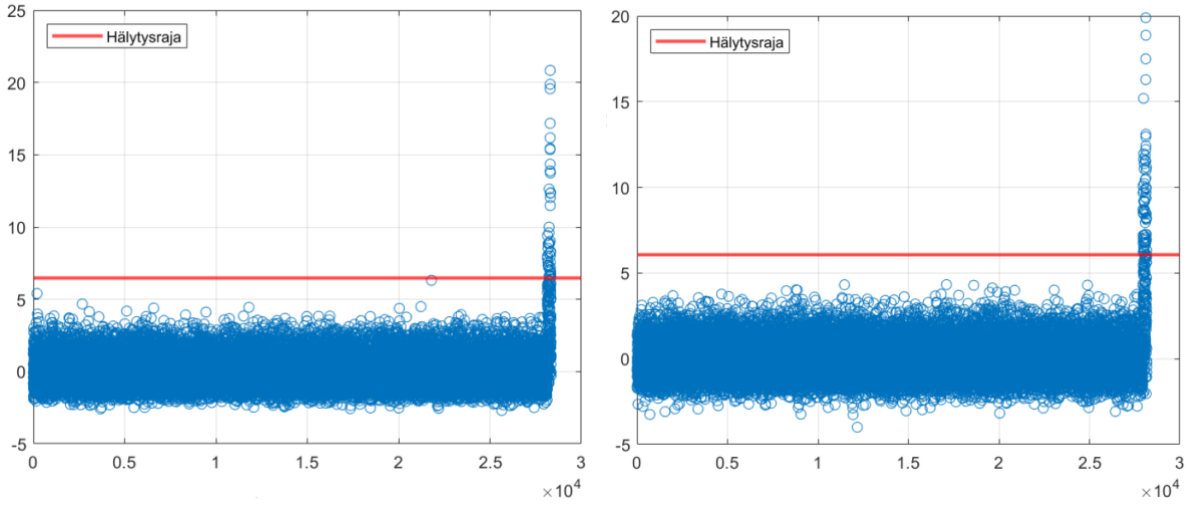

Figure 20. Left: Standardised residuals of the Nuorgam analysis data set. Right: Standardised residuals of the Rovaniemi analysis data set. In both cases the synthetic responses were added to the end of the dataset. The alarm limit is shown as a red line. Figures from (Vikman 2023).

4. Conclusions

The FFM method implemented for analysis of monitoring data sets is a viable method to detect anomalies in spectral time series. Synthetic spectra were detected in the data sets used to test the method.

The fixed features extracted from the training data, rain events and other contributions of environmental conditions, were taken into account by the nature of the algorithm, and false alarms were not produced in the tests.

A disadvantage of the method is that a training data set is needed prior to use. The training data set should contain all the features and behavior that is not due to artificial radioactivity. Thus, taking the algorithm to use requires that an already extensive data set is available.

Alternative ways to form the residuals used in deciding whether a measurement contains features not previously seen were studied only briefly. These methods were: the Mahalanobis distance and a modified Euclidean distance. In the modified Euclidean distance, the difference of an original channel content and the reconstruction only contributes to the residual if the former is higher than the latter. This method seemed to result in improved sensitivity for radionuclides that produce peaks, but reduced sensitivity for sources that produce continuums (such as x-rays). The usual Euclidean distance has the important advantage that the distribution of the residuals could be found. This allows the trade-off between false alarm rate and sensitivity to be controlled.

The algorithm could provide useful also for mobile data sets. Especially in a case where a background study is made of the area where radiation surveillance should be carried out. The underlaying assumptions, especially of the near normality of the channel contents, have to be fulfilled however.

5. References

Boardman, D., Reinhard, M., Flynn, A. 2012. Principal Component Analysis of Gamma-Ray Spectra for Radiation Portal Monitors; in IEEE Transactions on Nuclear Science, vol. 59, no. 1, pp. 154-160, Feb. 2012, https://ieeexplore.ieee.org/document/6138889

Eriksson, M, and Dowdall, M. 2021. On the use of principal component analysis, PCA, in gamma spectrometry NKS-B GAMMARAY X Webinar for users of gamma-ray spectrometry, 20-21 Oct 2021. Available at: http://halla.gr.is/wiki/GammaWiki/images/f/ff/Mats_Eriksson_On_the_use_of_PCA_in_gamma_spectrometry_.pdf

James, G., Witten, D., Hastie, T., Tibshirani, R. 2014. An Introduction to Statistical Learning: with Applications in R. Springer Publishing Company, Incorporated. 430 p.

Jolliffe, I. 2014. Principal Component Analysis. In Wiley Stats Ref: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat06472

Kishimoto, T., Woo, H., Komatsu, R., Tamura, Y., Tomita, H., Shimazoe, K., Yamashita, A., Asama, H. 2021. Path Planning for Localization of Radiation Sources Based on Principal Component Analysis. Appl. Sci. 2021, 11, 4707. https://doi.org/10.3390/app11104707

Kudomi, N. 1999. Energy calibration of plastic scintillators for low energy electrons by using Compton scatterings of c rays Nuclear Instruments and Methods in Physics Research A 430 (1999) 96}99. https://doi.org/10.1016/j.nima.2008.06.031

Minty, B., Hovgaard, J. 2002. Reducing noise in gamma-ray spectrometry using spectral component analysis, Exploration Geophysics, 33:3-4, 172-176, https://www.tandfonline.com/doi/abs/10.1071/EG02172

Pires de Lima, R., Marfurt, K. 2018. Principal component analysis and K-means analysis of airborne gamma-ray spectrometry surveys. 2277-2281. https://library.seg.org/doi/10.1190/segam2018-2996506.1

Williams, B. 2019. Applications of Principal Component Analysis for Gamma-Ray Spectroscopy with Position-Sensitive Semiconductor Detectors. Ph.D. Thesis, The University of Michigan, United States. Available at: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/149861/btwill_1.pdf?sequence=1

Reinhardt S. 2014. Full spectrum analysis in environmental monitoring. Radiat Prot Dosimetry. 160(4):311-7. https://doi: 10.1093/rpd/ncu144

Vikman, E. 2023. Pääkomponenttianalyysi poikkeamien havaitsemiseen gammaspektrometrisissä aikasarjoissa. Diploma thesis at the University of Tampere 2023, Submitted for publication.

Dowdall, M., Peräjärvi, K., Karhunen T., Ehrs, S., Espensen C., Jonsson G., Analysis of the Performance of LaBr3 Detectors for Fresh Fallout Response (PERLAD). NKS-B series final report NKS-453, Available at https://www.nks.org/en/nks_reports/view_document.htm?id=111010214697717

Bibliographic Data Sheet **NKS-481**

Title	Principle Component Analysis as Applied to Qualitative Analysis of Mobile Measurement and Monitoring Data Sets (PRICOMOB)
Author(s)	M. Dowdall[1], Tero Karhunen[2], Ellinoora Vikman[2], Mats Eriksson[3] Gísli Jónsson[4]
Affiliation(s)	[1]Norwegian Radiation Protection Authority, PO Box 55, N-1332, Østerås, Norway, [2]Radiation and Nuclear Safety Authority of Finland, Jokiniemenkuja 1, 01370 Vantaa, Finland, [3]Linköping University, SE-581 83 Linköping, Sweden, [4]Icelandic Radiation Safety Authority, Raudararstigur 10, 150 Reykjavik, Iceland.
ISBN	978-87-7893-577-9
Date	January 2024
Project	NKS-B / PRIMOCOB
No. of pages	31
No. of tables	5
No. of illustrations	20
No. of references	12

Abstract
max. 2000 characters

Mobile measurement systems are the backbone of most responses to cases of orphan sources. Conducting mobile measurement surveys, irrespective of the platform utilised, is a non-trivial task with respect to the nature of the data being accrued – large volumes of discrete, often highly variable, data points where the signal of interest may be weak, superimposed on a constantly fluctuating background and only present for a tiny proportion of the overall data set. Principal Component Analysis (PCA), one of the most popular multivariate statistical technique, is a flexible statistical procedure that allows for the summarizing of the information content in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed in order to observe trends, jumps, clusters and outliers. The PRICOMOB project focussed on the application of PCA to mobile measurement and stationary scanning data to assess its performance in identifying source signals from a number of isotopes superimposed on a variable background signal typical of mobile measurement data. The PCA method implemented proved itself a viable method to detect anomalies in spectral time series. A disadvantage of the method employed is that a training data set is needed containing all the features and

behavior that are not due to artificial radioactivity. Alternative ways to form the residuals used in deciding whether a measurement contains features not previously seen included the Mahalanobis distance and a modified Euclidean distance. The modified Euclidean distance seemed to result in improved sensitivity for radionuclides that produce peaks, but reduced sensitivity for sources that produce continuums (such as x-rays).

Key words Principal Component Analysis, Gamm spectrometry, Mobile measurement, time series

Available on request from the NKS Secretariat, P.O.Box 49, DK-4000 Roskilde, Denmark.
Phone (+45) 4677 4041, e-mail nks@nks.org, www.nks.org